TECHNICAL PAPER

ISSN 1047-3289 *J. Air & Waste Manage. Assoc.* **52**:1422–1432

Rapidly Locating and Characterizing Pollutant Releases in Buildings

Michael D. Sohn, Pamela Reynolds, Navtej Singh, and Ashok J. Gadgil
Indoor Environmental Department, Lawrence Berkeley National Laboratory, Berkeley, California

ABSTRACT

Releases of airborne contaminants in or near a building can lead to significant human exposures unless prompt response measures are taken. However, possible responses can include conflicting strategies, such as shutting the ventilation system off versus running it in a purge mode or having occupants evacuate versus sheltering in place. The proper choice depends in part on knowing the source locations, the amounts released, and the likely future dispersion routes of the pollutants. We present an approach that estimates this information in real time. It applies Bayesian statistics to interpret measurements of airborne pollutant concentrations from multiple sensors placed in the building and computes best estimates and uncertainties of the release conditions. The algorithm is fast, capable of continuously updating the estimates as measurements stream in from sensors. We demonstrate the approach using a hypothetical pollutant release in a five-room building. Unknowns to the interpretation algorithm include location, duration, and strength of the source, and some building and weather conditions. Two sensor sampling plans and three levels of data quality are examined. Data interpretation in all examples is rapid; however, locating and characterizing the source with high probability depends on the amount and quality of data and the sampling plan.

INTRODUCTION AND MOTIVATION

Airborne contaminant releases in or near a building can lead to significant human exposures unless prompt response measures are taken. However, possible responses can include conflicting strategies, such as shutting the ventilation system off versus running it in a purge mode or having occupants evacuate versus sheltering in place. The proper choice depends in part on knowing the source locations, the amounts released, and the likely future dispersion routes of the pollutants. Determining this information is complicated by the complex nature of airflows typically found in multi-room, multi-floor buildings. For example, merely detecting an airborne pollutant from sensors placed in the building may not reveal the location or strength of the source. The sensor measurements must be interpreted to estimate the source characteristics and quantify the uncertainties. For effective decision-making, the measurements must also be interpreted quickly and continuously as data stream in from the sensors.

Traditional algorithms for data interpretation generally use an inverse modeling approach (e.g., optimization and Gibbs sampling) to fit an indoor airflow and pollutant transport model to measurements of airborne pollutants. The fit is usually achieved by iteratively adjusting model input parameters until they reasonably predict the data. For online, real-time sensor data interpretation, these approaches are too slow. They (1) wait to execute computationally intensive fate and transport models until data are first obtained, (2) execute the models repeatedly as new or successive sensor data become available, and (3) require a considerable amount of data before the algorithm finds a unique solution or estimates the uncertainty in the calibrated parameters. Finally, the computational burdens required by these algorithms can be so great that using them for pre-event planning, such as to determine optimal monitoring locations, sampling plans, and sensor performance criteria, can be excessively cumbersome.

Many of these problems can be solved using a technique called Kalman filtering.[1] It is well-suited for many sensor interpretation applications and has been successfully applied, for example, to estimate the source strength of pollutant releases in multizone buildings.[2] However, Kalman filtering is best used for linear systems with well-conditioned input-to-output parameter covariance matrices and strong observability between the internal-state

IMPLICATIONS

This article presents and demonstrates an algorithm for interpreting multiple sensor measurements in real time. It may be used to characterize an unexpected pollutant release in or near a building. It may also be used to optimally place sensors, operate them, and determine their optimal performance criteria, including tradeoffs between sensitivity, reset time, and cost.

variables (e.g., the model input parameters of an indoor airflow and pollutant transport model) and the model outputs (e.g., concentration predictions).[1] Many fate and transport phenomena, such as second-order pollutant degradation, density-driven pollutant transport, aerosol coagulation, and second-order pollutant diffusion in sorption-desorption, are not linear. Furthermore, wide uncertainty bounds exist for several of the model inputs, such as many possible source locations, amounts released, durations of releases, and heating, ventilation, and air conditioning (HVAC) operating conditions. These will invariably lead to ill-conditioned covariance matrices and poorly observable systems. Although many nonlinear models may be linearized using an extended Kalman filtering technique,[3] the technique requires considerable tuning and adjustments because of the linear approximations.

Finally, Kalman filtering is not well-suited for the uncertainty analyses required for effective decision analysis. Kalman filtering may be applied to estimate the most likely source location and uncertainty (e.g., "The source is in room 'A' with 80% probability"). It is, however, considerably more computationally intensive, and thus time-consuming, to concurrently estimate the uncertainties of other less likely source locations (e.g., "The source is in room 'A' with 80% probability, in room 'B' with 18% probability, and in room 'C' with 2% probability").

We present an alternative algorithm that uses Bayesian statistics. This approach succeeds where traditional methods fail because it decouples the simulation of predictive fate and transport models from the interpretation of measurements, and incorporates uncertainty analysis in all parts of the framework. Thus, we can compute the time-consuming airflow and pollutant transport predictions and uncertainty estimates—without requirements on linearity of the models—before a pollutant release event and interpret sensor data in real time during an event. The technique may be used to estimate the location, magnitude, and duration of the release, to characterize any unknown or variable building or weather conditions, and to predict future pollutant transport in the building. Initial estimates are provided as soon as a sensor detects a pollutant and can be updated as each new measurement arrives.

The techniques we introduce are not new; often termed "Bayes Monte Carlo updating," it has been applied to assess environmental health risk,[4–6] analyze groundwater monitoring data,[7–9] and conduct environmental value-of-information analyses.[10,11] However, the research problems described in these articles are distinct from the current work in one important feature. They describe applications to interpret data well after they were collected, when interpretation and response were not time-critical. In the present work, we exploit a feature of Bayes Monte Carlo updating that has not, to our knowledge, been previously recognized: Modeling and data analysis can be decoupled, which allows for data to be interpreted while they stream in during a pollutant release event. This is a significant advance over previous uses of Bayes Monte Carlo methods. Furthermore, application of this general approach to indoor air pollutant source characterization and airborne pollutant transport predictions has not, to our knowledge, been reported in the literature.

The objectives of this article are thus to present a Bayesian algorithm for interpreting sensor data in real time, and demonstrate the approach by successfully detecting and characterizing a pollutant release in a hypothetical five-room building. In the illustrative application, we generated synthetic data for two data collection scenarios: (1) concurrent sampling, in which sensor measurements are obtained simultaneously in each of the five rooms at 5-min intervals; and (2) sequential sampling, in which sensor measurements are obtained sequentially, one room at a time, at 5-min intervals. We also generated high-, medium-, and low-quality data for each of the two data collection scenarios to examine degradation of the predictive results with increasingly noisy data. In addition to unknowns regarding the location, duration, and magnitude of the pollutant release, other unknowns included the outside temperature and whether certain doors or windows were open or closed. Finally, interpreting data for several sampling plans and qualities of data demonstrates the ease of exploring and comparing the tradeoffs among sensor features, such as frequency of sampling, sensor sensitivity, and number of sensors.

APPROACH

The Bayesian data interpretation approach is divided into two stages. First, in the pre-event or simulation stage, the practitioner selects a fate and transport model, builds a computer model of the building, characterizes uncertainties of the model inputs, and simulates many hypothetical airflow and pollutant transport scenarios. These time-consuming tasks are completed before a pollutant release occurs. In the second stage, during a pollutant release event, the agreement between each of the model simulations and sensor data is evaluated using a technique called Bayesian updating.[5,9] This stage is quick and is conducted as data stream in from the sensors.

Pre-Event Planning

Before a release event, the practitioner develops a model of the building's indoor airflow and pollutant transport. Best estimates for model inputs are generated from, for example, previous building characterization exercises,

tracer gas flow experiments and modeling, published literature, and professional judgment. Any uncertain model input parameter or variable building characteristic is assigned an uncertainty distribution that describes the probabilistic range of possible values. Pollutant description uncertainties, such as the location, duration, and amount of pollutant released in an incident, are also assigned uncertainty distributions. In general, wide distributions are assigned because of limited prior information, particularly for describing the pollutant characteristics.

The practitioner next generates a library of model simulations by sampling the distributions of the model input parameters using a Monte Carlo or other sampling technique and predicting airflow and pollutant transport for each set of parameters. Each model simulation represents a possible building configuration and pollutant release scenario. At this stage, each simulation is equally likely to occur. Thus, sufficient sampling of the uncertainty distributions is essential to represent the full range of possible building and pollutant release characteristics. One method for testing sufficiency of sampling is by increasing the sample size until changes in summary statistics (e.g., means, variances, coefficients of variation) of model predictions are negligible. The resulting library of simulations may consist of several thousand scenarios. Because this stage is not time-critical, a large library of simulations is not difficult to develop with the advances of fast personal computers and inexpensive data storage devices.

It is important that the parameter ranges sampled in the uncertainty characterization and Monte Carlo sampling are wide enough to contain the parameter values of the actual event (to be diagnosed in real time). Otherwise, the method will fail to converge on the correct parameter values. Of course, such failure still provides some useful information; for example, the actual event parameters are not within the ranges sampled, or there is a model-misfit to the actual event.

During Event Data Interpretation

During a release event, the algorithm compares data streaming in from sensors to each realization in the library of model simulations using a structured probabilistic method referred to as Bayesian updating. Bayes' rule allows the practitioner to quickly estimate and update the level of agreement between the observed data and model simulations (i.e., the pollutant transport predictions). To summarize the process, the practitioner compares each realization in the library to the data to assess the likelihood that the realization describes the event in progress. A model simulation with predictions that fit the sensor data well will have a high likelihood estimate. This in turn suggests that the model inputs used to generate that realization in the pre-event simulation stage have a high probability of describing the event in progress. By comparing the relative fits for each realization using Bayesian statistics, the practitioner estimates the best-fitting suite of model inputs and the associated uncertainty.

The difference between the data and the predictions resulting from measurement error, spatial and temporal averaging or correlations, and imperfect model representation are all considered when estimating the data-to-model agreement. The probability of each model simulation before and after assessing the agreement is termed the prior and posterior probability, respectively.

The posterior probability of the kth Monte Carlo simulation making prediction Y_k given the sensor measurements O is denoted as $p(Y_k|O)$. Using Bayes' rule, $p(Y_k|O)$ is calculated using eq 1[5]

$$p(Y_k|O) = \frac{L(O|Y_k)p(Y_k)}{\sum_{i=1}^{K} L(O|Y_i)p(Y_i)} \quad (1)$$

where $p(Y_k|O)$ is the posterior probability, $L(O|Y_k)$ is the likelihood of observing measurements O given model prediction Y_k, $p(Y_k)$ is the prior probability of the kth Monte Carlo simulation, and K is the number of Monte Carlo simulations. Before data comparison, each of the model realizations is usually assumed to be equally likely (i.e., $p(Y_k) = 1/K$).

The posterior probability, $p(Y_k|O)$, describes the probability of all of the model assumptions and predictions associated with the kth realization. Thus, the prior uncertainty of each model input parameter (e.g., source location or building characteristic) and model output (e.g., airborne concentration prediction) is updated according to how well model predictions in the prior uncertainty distribution agree with the sensor data. The updated mean, variance, and correlation coefficient of each model input parameter and output are calculated using eqs 2–4, respectively[5]

$$\mu'_V = \sum_{i=1}^{K} V_i \cdot p(Y_i|O) \quad (2)$$

$$\sigma'^2_V = \sum_{i=1}^{K} (V_i - \mu'_V)^2 \cdot p(Y_i|O) \quad (3)$$

$$\rho'_{V,W} = \frac{\sum_{i=1}^{K} (V_i - \mu'_V)(W_i - \mu'_W) \cdot p(Y_i|O)}{\sigma'_V \cdot \sigma'_W} \quad (4)$$

where V and W represent any model input or output.

The likelihood function, $L(O|Y_k)$, in eq 1 quantifies the error structure of the data—that is, the differences between the data and the model predictions resulting from measurement error, spatial and temporal averaging or correlations, and imperfect model representation. If many independent measurements are considered, for example, following sequential concentration measurements returned from sensors or from concurrent measurements sampled in several locations, the likelihood of observing all of the measurements is the product of all of the individual likelihoods

$$L(O|Y_k) = \prod_{s=1}^{S} L(O_s|Y_{s,k}) \quad (5)$$

where S is the number of independent measurements.

For unbiased measurements with a normally distributed error, the likelihood of observing a sensor measurement, O_s, given a model prediction, $Y_{s,k}$, is given as

$$L(O_s|Y_{s,k}) = f(O_s - Y_{s,k}) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{O_s - Y_{s,k}}{\sigma_\varepsilon}\right]^2\right) \quad (6)$$

where O_s is the concentration measured by a sensor in a room at $t = t_s$, $Y_{s,k}$ is the airborne concentration predicted from the kth Monte Carlo realization that corresponds to O_s, and σ_ε^2 is the error variance of the measurements. The error variance, σ_ε^2, describes not only the error in the sensor instruments but also the error associated with comparing model predictions with sensor measurements having different spatial and temporal averaging.

Although a Gaussian likelihood function, appropriate for independent normally distributed errors, is commonly applied by researchers in various environmental applications as in eqs 5 and 6,[4,5,9,11] it assumes an error structure that is inappropriate when errors in the data are correlated.[9] For example, the sensors may have a calibration bias that causes all of the measurements to under- or over-report the concentrations. Sohn et al. discuss alternative methods for estimating likelihood functions in these cases.[9] However, for the purposes of our illustrative application, we assumed that the measurement errors were uncorrelated and could be described by a Gaussian likelihood function, although alternatives can be readily implemented.

The second stage of the approach is mathematically simple and can be executed very quickly, much more quickly than the rate at which new data are likely to arrive from sensors.

ILLUSTRATIVE APPLICATION

We applied our approach to locate and characterize a hypothetical pollutant release in a five-room building. Uncertainties in source location, duration, and amount, and in some building characteristics, were estimated and updated using synthetic data.

The subsections of this article describing the preevent planning consist of a description of the five-room building, the uncertainty characterization of model input parameters, and the airflow and pollutant transport predictions. A description of the synthetic data follows. Data interpretation during an event is described next. Then, several examples of rapid data interpretation are discussed for various scenarios of data sampling and data quality.

Building Description

The study building is a single-story building comprising three rooms, a common area (CA), and a bathroom (Figure 1). Each of the partitioned areas is treated as a well-mixed zone. The zones connect to the outside via windows and doors and interconnect via internal doors. The building does not have an HVAC system.

The interior door between the CA zone and Room 1 is open, as are the windows in the bathroom and Room 2. The status of one of the CA zone windows and the door between the CA zone and Room 3 is unknown (for example, owing to failed position sensors at these locations). These are denoted in Figure 1 with question marks. All other windows and doors are closed. Wind blows at a steady 3 m/sec on the exterior wall shared by the CA zone and Room 1. The temperatures of the rooms are indicated in Figure 1, and the outside temperature is unknown (see Table 1).

Table 1 summarizes the uncertainties in the source and building characteristics. Although the uncertainty

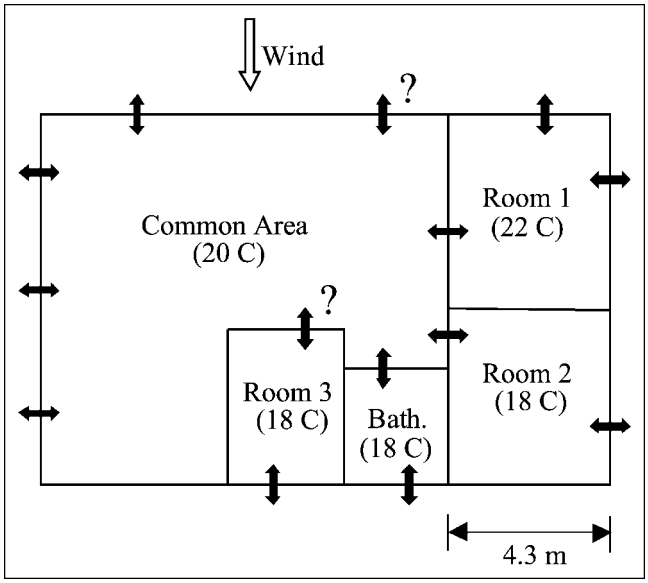

Figure 1. Plan of the five-room building. The arrows represent windows or doors; the question marks indicate uncertain open or closed status. The windows in the bathroom and Room 2 and the interior door for Room 1 are open. All other windows and doors are closed. The wind blows at a steady 3 m/sec.

Table 1. Uncertainty in the source and building characteristics before data interpretation.

Parameter	Range	Distribution
Source location	Any room	Equiprobable
Total mass released	5–100 g	Uniform
Release duration	5–20 min	Uniform
Outside temperature	10–25 °C	Uniform
Room 3 door position[a]	Open or closed	Equiprobable
Common area window position[a]	Open or closed	Equiprobable

[a]The door and window are identified on Figure 1.

distributions were assumed to be wide, we did not presume to capture the full extent of the building variability. For this illustrative application, the uncertainties were limited to only some aspects of the pollutant and building characteristics. Nevertheless, they demonstrate the types of uncertainty that can be estimated simultaneously from the data.

Airflow and Pollutant Transport Simulation

We selected the COMIS model to predict indoor airflow and pollutant transport.[12] COMIS predicts the steady-state flows of air and the dynamic transport of pollutants by representing the building as a collection of well-mixed zones. Air flows between zones via cracks, doors, and windows (and also fans and ductwork, although those features were not used here). COMIS assumes air to be incompressible and calculates airflow through these pathways based on pressure differences across them, induced by wind and thermal buoyancy. The model calculates dynamic pollutant transport assuming that the pollutant transports at the same rate as air. COMIS has been used to predict airflow and gas transport in multi-story, low- and high-rise residences,[13,14] small office buildings,[15] controlled experimental test houses,[16] and single-family houses.[17] Although we selected a multizone modeling approach for this application, our data interpretation algorithm may be used with any suitable indoor airflow and pollutant transport model.

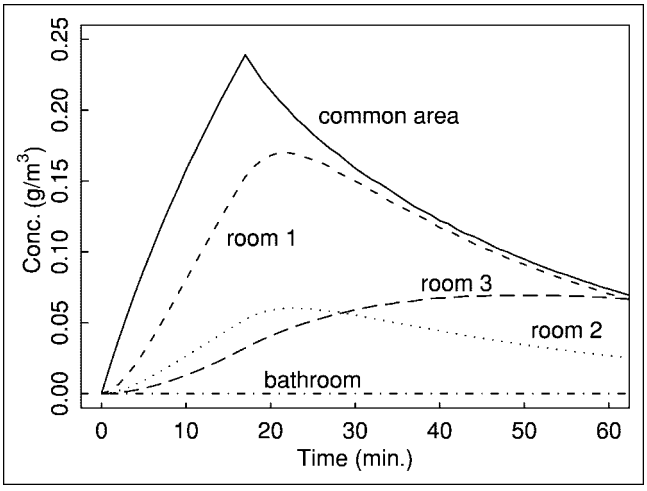

Figure 2. Simulation of a 57-g, 17-min release in the CA zone (steady release rate). The outside temperature is 12.7 °C. The interior temperatures are shown in Figure 1. The windows in the bathroom and Room 2 and the Room 1 interior door are open. All other windows and doors are closed.

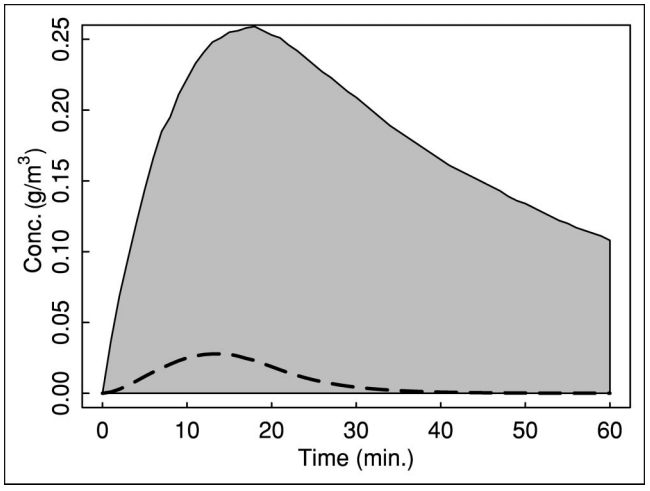

Figure 3. Concentration prediction in the CA zone before data interpretation. The gray area is the 90% confidence interval from 5000 simulations of pollutant releases and building characteristics. The dotted line is the median prediction.

As part of the pre-event planning, we generated a library of airflow and pollutant transport simulations by sampling the uncertainty distributions describing the source and building characteristics (Table 1). Five thousand air flow and pollutant simulations, each of them equally likely, were generated using Latin Hypercube sampling techniques.[18] Means and variances for several sample sizes were tested to ensure that 5000 simulations adequately sampled the problem solution space.

Figure 2 shows an example of one of the simulations. A pollutant was released in the CA zone for 17 min at 3.4 g/min. The high concentrations in Room 1 result from the high airflow through the open door to the CA zone. The lower concentrations in Rooms 2 and 3 and the bathroom result from closed interior doors and infiltration/exfiltration with the outside. Figure 3 shows the

Table 2. Two-part error structure used to generate the synthetic data.

Data Quality	Coefficient of Variation	Random Error, σ (g/m^3)
High	0.05	0.005
Medium	0.1	0.01
Low	0.5	0.05

Table 3. Updated source and building uncertainties. The Answer row lists the values used to generate the synthetic data. The Prior row is before any data interpretation. Concurrent sampling draws a measurement from all rooms every five minutes. Sequential sampling draws a measurement sequentially one room at a time in the order: (1) CA at t = 5 min. (2) Room 1 at t = 10 min. (3) Room 2 at t = 15 min. (4) bathroom at t = 20 min. (5) Room 3 at t = 25 min. and (6) CA at t = 30 min.

Time (min)	Data Quality	Total Mass (g)		Release Duration (min)		External Temp. (°C)		Probability of CA Window Closed	Probability of Room 3 Door Closed
		μ	σ	μ	σ	μ	σ		
Answer	—	57	—	17	—	12.7	—	100	100
Prior	—	53	27	12.5	4.3	17.5	4.3	50	50
Concurrent Sampling									
5	high	47	14	13.2	3.9	14.3	2.9	100	100
	medium	40	13	13.6	4.3	15.4	2.9	100	100
	low	34	23	14.0	4.0	15.1	2.9	42	53
10	high	54	10	15.1	2.9	13.0	2.6	100	100
	medium	46	10	14.8	3.1	14.9	3.1	100	100
	low	32	18	14.0	4.1	15.3	2.9	49	57
15	high	56	8	15.6	2.5	12.2	2.1	100	100
	medium	47	7	15.6	2.6	13.4	2.8	100	100
	low	41	16	14.4	3.7	15.4	2.9	80	73
20	high	59	7	17.0	0.1	11.9	1.9	100	100
	medium	53	3	18.4	1.4	14.0	2.5	100	100
	low	47	13	15.0	3.3	15.4	3.0	97	92
25	high	59	0.2	17.0	0.1	11.9	0.1	100	100
	medium	54	1	19.3	0.9	14.1	1.6	100	100
	low	49	10	15.5	3.1	15.8	2.9	100	100
30	high	59	0.3	17.0	0.1	11.9	0.1	100	100
	medium	53	1	18.6	1.0	12.8	1.3	100	100
	low	47	9	15.0	3.2	15.5	2.9	100	100
Sequential Sampling									
5	high	67	20	11.0	4.2	15.3	2.8	63	44
	medium	63	22	11.6	4.2	15.1	2.8	66	43
	low	52	27	12.7	4.3	15.0	2.9	51	48
10	high	67	18	12.5	4.3	15.1	2.9	51	46
	medium	59	21	12.5	4.3	15.0	2.9	59	40
	low	48	27	13.0	4.3	15.0	2.9	50	47
15	high	57	13	12.1	4.4	16.0	2.8	80	35
	medium	57	21	12.2	4.3	15.0	2.8	77	40
	low	48	26	12.8	4.4	15.1	2.9	49	44
20	high	56	13	13.8	4.0	16.5	2.6	100	38
	medium	53	20	13.8	4.0	14.9	2.8	91	34
	low	43	25	12.2	4.3	14.9	2.9	45	44
25	high	57	8	15.4	2.0	13.1	2.1	100	100
	medium	46	10	15.7	3.7	13.3	2.3	97	100
	low	36	24	11.6	4.2	14.9	2.9	30	47
30	high	59	4	16.7	1.0	13.1	1.7	100	100
	medium	46	5	16.5	1.8	12.5	1.8	100	100
	low	35	23	11.6	4.2	14.9	2.9	40	50

confidence interval for the predicted pollutant levels in the CA zone based on all of the realizations. Not knowing the pollutant release characteristics or some of the building conditions results in highly uncertain airborne concentration predictions. The uncertainty bounds will be reduced as data are interpreted during an event.

Description of Synthetic Data

We generated synthetic data to represent measurements that might stream in from air monitoring sensors placed in the building. The synthetic data were based on an airflow and pollutant transport simulation representing a possible pollutant release event. Figure 2 plots the

simulation from which the synthetic data were generated, and Table 2 summarizes the model input information. The simulation was excluded from the library of 5000 simulations describing the prior pollutant concentration predictions.

We added measurement error to the model simulation using a two-part error structure: (1) a normally distributed error associated with a standard deviation proportional to the true value—that is, a fixed coefficient of variation, and (2) a normally distributed random error independent of the magnitude of the measurement. Part one of the error structure represents error associated with the magnitude of the airborne concentration, and part two represents error caused by random noise. Thus, the error variance in eq 6 is equal to the algebraic sum of the two-part errors at each sampling time. More complex errors, such as lognormal error structures, temporally or spatially correlated measurement errors, or errors caused by incomplete mixing in the room, were not used, though they can often occur. Many statistical methods for handling these error structures are available and can be used in place of eq 6.[9,19]

We generated high-, medium-, and low-quality synthetic data with progressively larger magnitudes of error components in the error structure. If adding the error generated a negative value for the pollutant concentration, the simulated measurement was set to zero. Table 3 summarizes the error components for the three levels of data quality, and Figure 4 shows the synthetic data in the CA zone. As expected, the high-quality data (Figure 4a) show a more consistent pollutant concentration time series than do the low-quality data (Figure 4c).

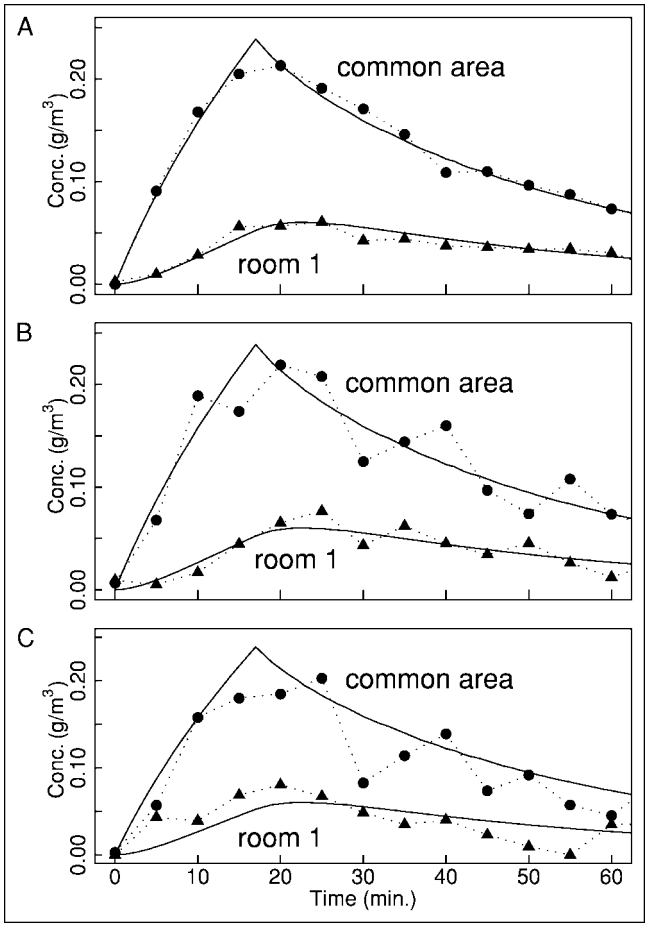

Figure 4. Synthetic data representing (a) high-, (b) medium-, and (c) low-quality measurements in the CA zone and Room 1. The solid lines represent the model simulation originally used to generate the data.

Data Interpretation

The data interpretation algorithm presented in the "Approach" section allowed us to easily evaluate alternative data gathering plans without re-running the computationally intensive fate and transport models. Thus, along with the three levels of data quality (described in the synthetic data section), we also evaluated two different plans for data collection. In the first, concurrent sampling, we obtained sensor measurements from all five zones simultaneously at 5-min intervals. In the second, sequential sampling, we obtained sensor measurements sequentially, one zone at a time, at 5-min intervals. In the sequential sampling plan, we also examined whether the sequence of the sampling affected the data interpretation results. In each application, data interpretation was rapid and was conducted while data streamed in from the sensors.

Although we present the results for the various combinations of data quality and sampling plans, it is important to emphasize that the results merely illustrate the types of data interpretation and "what-if" analyses that may be conducted using our interpretation algorithm. The results do not represent the success or failure of the interpretation approach. They reveal some of the tradeoffs between benefits and costs that an end user must consider when deploying and operating a sensor network.

Data Interpretation Using the Concurrent Sampling Plan. Figure 5 shows the estimation of the source location for the three qualities of data. With medium- or high-quality data, the interpretation correctly identifies the source location at the first measurement event (t = 5 min), when five measurements were obtained. With low-quality data, the identification of the source location is slower, requiring more measurements to overcome the error in the data. Table 2 summarizes the results at several sampling times. Again, the medium- and high-quality data permit dramatic uncertainty reductions at t = 5 min, in all cases converging to the correct answers. The low-quality data, however, required more data and thus more time.

Data Interpretation Using the Sequential Sampling Plan. Next, we performed the same estimation with data obtained using a sequential sampling plan. This plan

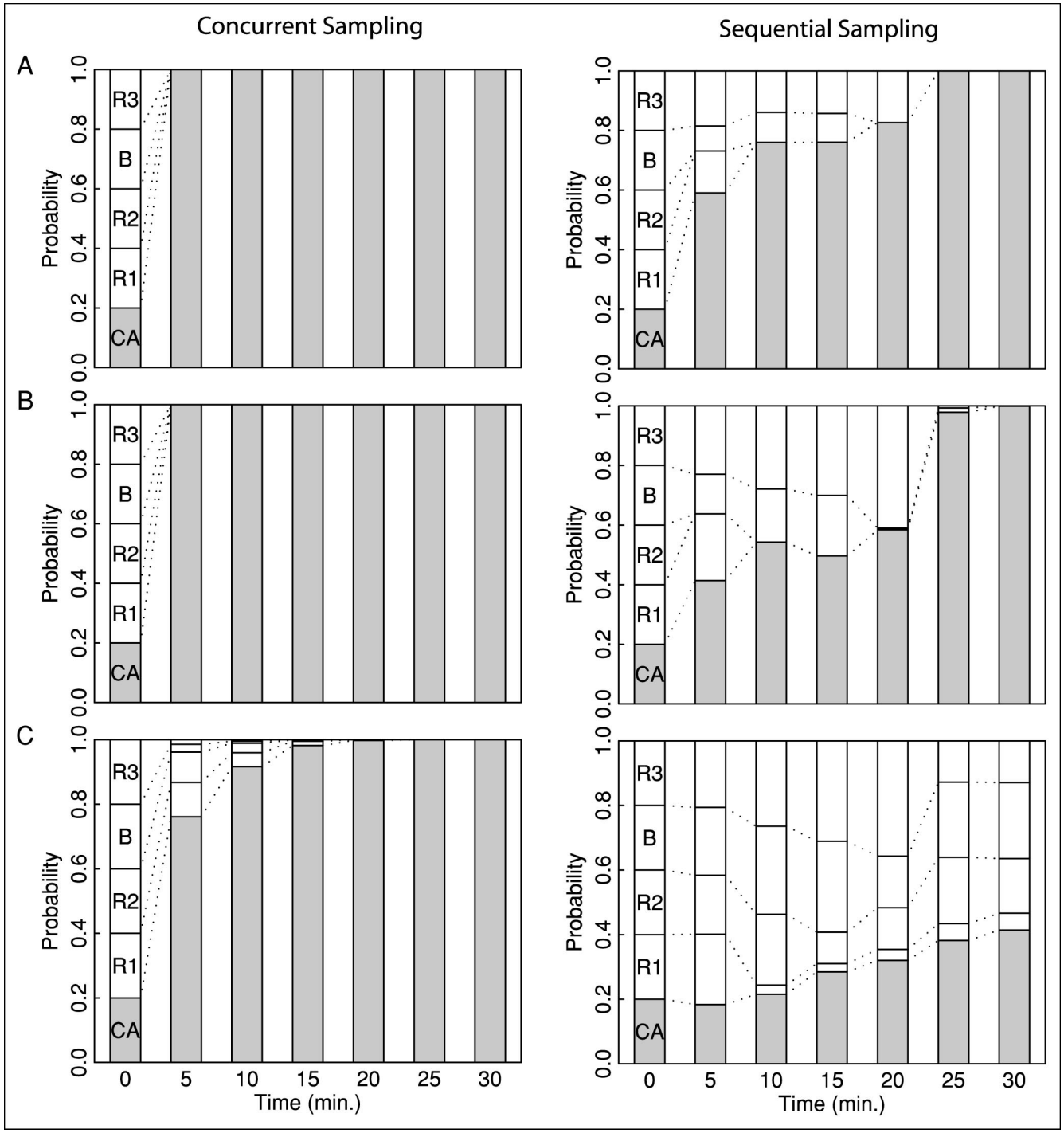

Figure 5. Locating the source using (a) high-, (b) medium-, and (c) low-quality measurements. Concurrent sampling draws a measurement from each zone every 5 min. Sequential sampling draws a measurement from one room at a time every 5 min in the order given in the text. The probability at t = 0 is before data interpretation.

gathers data at a significantly slower pace—one data point every 5 min instead of five. Sequential sampling can represent a situation where a single (expensive) sensor is multiplexed to several sampling tubes. The rooms were sampled in this order: (1) CA zone at t = 5 min, (2) Room 1 at t = 10 min, (3) Room 2 at t = 15 min, (4) bathroom at t = 20 min, (5) Room 3 at t = 25 min, and (6) CA zone at t = 30 min. Figure 5 shows the estimation of the source location. Bearing in mind that the sequential sampling plan collects five times less data than the concurrent sampling plan, the medium- and high-quality data do not locate the source until all of the rooms are sampled once (t = 25 min), although reasonably good estimates are generated as early as t = 10 min. The low-quality data case, however, does not locate the source even after 30 min.

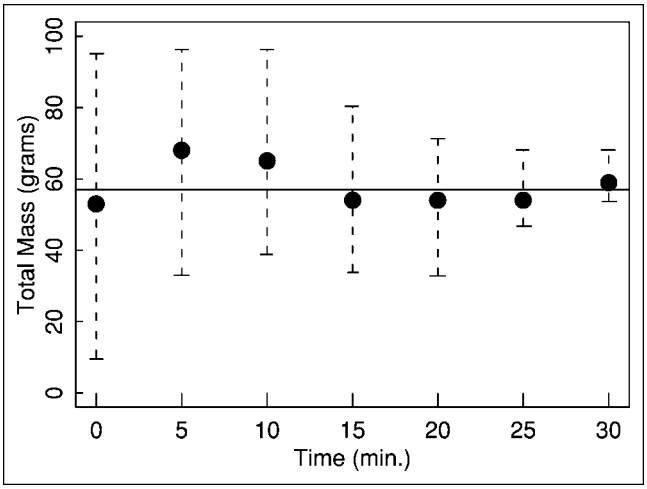

Figure 6. Estimating the total mass released by sequentially sampling and obtaining high-quality measurements. Sequential sampling was in the order given in the text. The uncertainty tails represent the 90% confidence interval, and the circle represents the median prediction. The horizontal line is the value originally used to generate the synthetic data.

Figure 5 also shows that the interpretation algorithm predicts location probabilities of rooms not yet sampled. For the medium- and high-quality data, at t = 5 min, when the CA zone is sampled, the probability of Room 2 is nearly zero. The pre-event library contained few simulations where air flowed from Room 2 to the CA zone because of the wind direction and temperature differences between rooms and between the indoors and outdoors. Thus, a pollutant released in Room 2 would not transport to any other room. Because a nonzero concentration is measured in the CA zone at t = 5 min, the source must not be located in the bathroom.

Figure 6 shows the estimated release amount for the high-quality data set. The gradual reduction of uncertainty was consistent with the 5-fold-less data obtained at each time step. Table 2 summarizes the estimation of all of the pollutant and building characteristics. Though not plotted, the results showing successive improvements in the estimates were consistent with the successive uncertainty reductions illustrated in the source location estimates (Figure 5) and source amount estimates (Figure 6).

Finally, we again performed the estimation with data obtained using a different sequential sampling plan. The rooms were sampled in this order: (1) Room 2 at t = 5 min, (2) bathroom at t = 10 min, (3) Room 3 at t = 15 min, (4) CA zone at t = 20 min, (5) Room 1 at t = 25 min, and (6) Room 2 at t = 30 min. The CA zone, where the source is located, is sampled much later in this sequence.

Figure 7 shows the estimation of the source location under this sampling plan. With the high-quality data, the source is located after all of the rooms are sampled (t = 25 min), as in the previous sampling sequence. However, the medium-quality data set poses greater difficulties than before, and the interpretation of the low-quality data again fails. This result further supports the importance of a data interpretation algorithm that is capable of easily examining alternative sampling plans and qualities of data. These capabilities are essential for properly designing a sampling plan that balances the effectiveness of

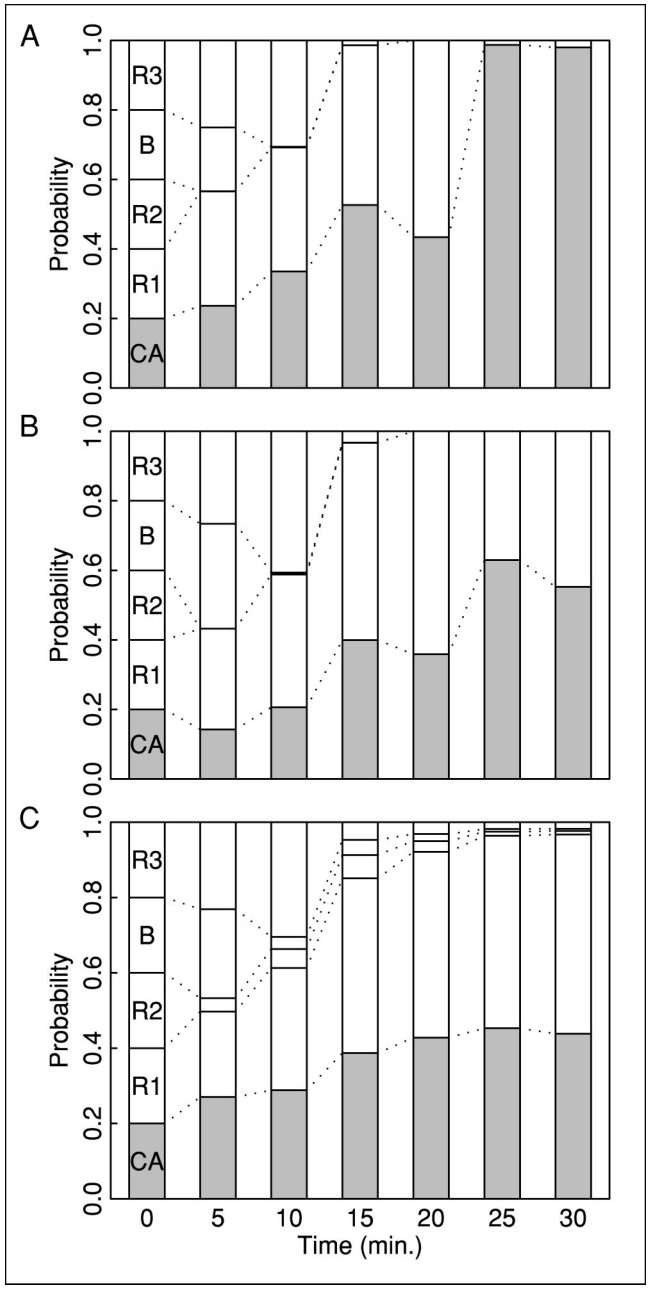

Figure 7. Locating the source using (a) high-, (b) medium-, and (c) low-quality measurements. Sequential sampling was in the order given in the text.

deploying a monitoring system in a building with the costs of operating it or selecting the appropriate equipment.

Estimating Pollutant Transport. Along with estimating pollutant and building characteristics, the data interpretation algorithm also can predict future transport behavior of the pollutant. Recall that V and W in eqs 2–4 can represent any input parameter, such as pollutant and building characteristics, or any output, such as pollutant concentration. Figure 8 shows the predicted concentration profile for the CA zone as high-quality data sequentially streams in from the sensors using (1) the CA zone, (2) Room 1, (3) Room 2, (4) the bathroom, (5) Room 3, and (6) the CA zone sequence. The concentration predictions before data interpretation—that is, from the equally weighted original 5000 simulations—are shown in Figure 2. In Figure 8, at each step of the sampling sequence, the data interpretation algorithm updates the best guesses and the uncertainty. Similar to the results for estimating the source location, the concentration predictions do not show significantly reduced uncertainties until the CA zone is sampled again at t = 30 min. Nevertheless, it is remarkable that the posterior median concentration estimate is already very close to the correct answer at t = 5 min, when the CA zone is first sampled. The concurrent sampling plan dramatically reduced uncertainty at t = 5 min and, therefore, is not shown.

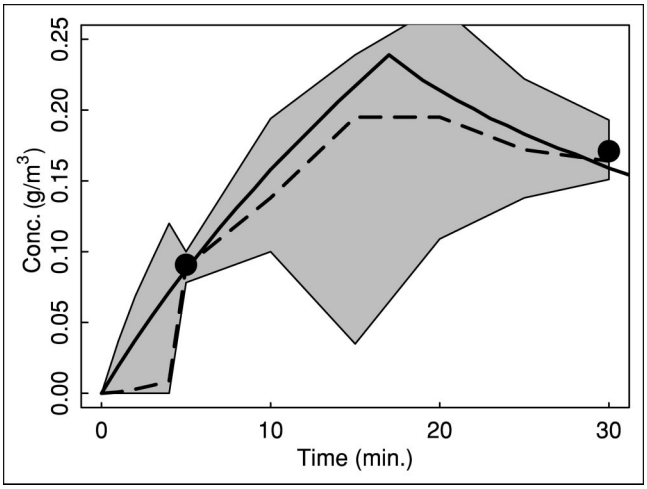

Figure 8. Predicting pollutant concentrations in the CA zone using sequential sampling and obtaining high-quality measurements. Sequential sampling was in the order given in the text. The gray area shows the 90% confidence interval, and the dotted line is the median. The circles represent synthetic data gathered in the CA zone. The solid line depicts the model simulation originally used to generate the data. Figure 3 shows the concentration predictions in the common area before data interpretation.

CONCLUSION

This article presents a Bayes Monte Carlo approach for interpreting sensor measurements in real time. It may be used with sensors to characterize an unexpected pollutant release in or near a building. It may also be used to optimally place sensors, operate them, and determine optimal tradeoffs among their performance criteria.

Our data interpretation approach differs from previous work relating to model parameter estimation by decoupling the simulation of airflow and pollutant transport from the interpretation of measurements. This allows us to divide the analysis into two parts. The pre-event planning stage completes the time-consuming tasks such as development of airflow and pollutant transport models, building description, uncertainty characterization, and simulation of pollutant transport, and compiles the scenario simulations into a library of results. The data interpretation stage is then quickly executed, accessing this library as data stream in during a pollutant release event. The decoupled nature of this approach also allows easy and quick evaluation of alternative sampling plans or the performance of sensors without re-executing the time-consuming pre-event stage of the analysis.

We demonstrated the approach by analyzing a hypothetical pollutant release in a five-room building. Data interpretation estimated the location, total amount, and release duration of a pollutant, some building conditions, and future pollutant transport for three different plans for sensor sampling and three different qualities of measurement data. For each sampling plan, data interpretation was rapid and was conducted while data hypothetically streamed in from sensors; however, locating and characterizing the source with high probability depended on the amount and quality of data available and the sampling plan.

In future work, we will use our approach to guide sensor deployment. Decoupling data interpretation from model evaluation allows a comparison of the performance of many hypothetical sensor operating conditions and sensor locations. Such comparisons could help identify the requirements for a sensor network, including the number, sensitivity, and response time of sensors, based on the desired performance of a data interpretation algorithm in any given building.

ACKNOWLEDGMENTS

This work was supported by the Office of Non-Proliferation and National Security, Chemical and Biological Non-Proliferation Program, of the U.S. Department of Energy (DOE) under Contract No. DE-AC03–76SF00098. Reynolds was supported in part by the DOE Energy Research Undergraduate Laboratory Fellowship Program. Singh was supported in part by DOE under the Pre-Service

Teacher Program. We thank P.N. Price for assisting in the conceptual development and application of the Bayes Monte Carlo approach used in this research. We thank D.M. Lorenzetti, P.N. Price, E.A. Derby, and W. Davis for comments on an earlier version of the manuscript.

REFERENCES

1. Grewal, M.S.; Andrews, A.P. *Kalman Filtering: Theory and Practice,* 2nd ed.; Wiley & Sons: New York, 2001.
2. Federspiel, C.C. Estimating the Inputs of Gas Transport Processes in Buildings; *IEEE Transact.* **1997**, *5* (5), 480–489.
3. Del Gobbo, D.; Napolitano, M.; Famouri, P.; Innocenti, M. Experimental Application of Extended Kalman Filtering for Sensor Validation; *IEEE Trans. on Control Sys.; Technol.* **2001**, *9* (2), 376–380.
4. Taylor, A.C.; Evans, J.S.; McKone, T.E. The Value of Animal Test Information in Environmental Control Decisions; *Risk Anal.* **1993**, *13* (4), 403–412.
5. Brand, K.P.; Small, M.J. Updating Uncertainty in an Integrated Risk Assessment: Conceptual Framework and Methods; *Risk Anal.* **1995**, *15* (6), 719–731.
6. Pinsky, P.F.; Lorber, M.N. A Model to Evaluate Past Exposure to 2, 3, 7, 8,-TCDD; *J. Exp. Anal. Env. Epid.* **1998**, *8* (2), 187–206.
7. Dilks, D.W.; Canale, R.P.; Meier, P.G. Development of Bayesian Monte-Carlo Techniques for Water Quality Model Uncertainty; *Ecolog. Model.* **1992**, *62*, 149–162.
8. Wolfson, L.J.; Kadane, J.B.; Small, M.J. Bayesian Environmental Policy Decisions: Two Case Studies; *Ecolog. Applic.* **1996**, *6* (4), 1056–1066.
9. Sohn, M.D.; Small, M.J.; Pantazidou, M. Reducing Uncertainty in Site Characterization Using Bayes Monte Carlo Methods; *J. Env. Eng.* **2000**, *126* (10), 893–902.
10. Finkel, A.M.; Evans, J.S. Evaluating the Benefits of Uncertainty Reduction in Environmental Health Risk Assessment; *J. Air Pollut. Control Assoc.* **1987**, *37*, 1164–1171.
11. Dakins, M.E.; Toll, J.E.; Small, M.J.; Brand, K.P. Risk-Based Environmental Remediation: Bayesian Monte Carlo Analysis and the Expected Value of Sample Information; *Risk Anal.* **1996**, *16* (1), 67–79.
12. Feustel, H.E. COMIS—An International Multizone Air-Flow and Contaminant Transport Model; *Energy Build.* **1999**, *30*, 3–18.
13. Feustel, H.E.; Zuercher, C.H.; Diamond, R.; Dickinson, B.; Grimsrud, D.; Lipschutz, R. Temperature- and Wind-Induced Air Flow Patterns in a Staircase. Computer Modelling and Experimental Verification; *Energy and Build.* **1985**, *8*, 105–122.
14. Sextro, R.G.; Daisey, J.M.; Feustel, H.E.; Dickerhoff, D.J.; Jump, C. Comparison of Modeled and Measured Tracer Gas Concentrations in a Multizone Building. Presented at 8th International Conference on Indoor Air Quality and Climate, Indoor Air 1999; Paper Number 4–785.
15. Feustel, H.E. Measurements of Air Permeability in Multizone Buildings; *Energy Build.* **1990**, *14*, 103–116.
16. Haghighat, F.; Megri, A.C. A Comprehensive Validation of Two Airflow Models—COMIS and CONTAM; *Indoor Air* **1996**, *6*, 278–288.
17. Zhao, Y.; Yoshino, H.; Okuyama, H. Evaluation of the COMIS Model by Comparing Simulation and Measurement of Airflow and Pollutant Concentration; *Indoor Air* **1998**, *8*, 123–130.
18. Iman, R.L.; Conover, W.J. Small Sample Sensitivity Analysis Techniques for Computer Models, with an Application to Risk Assessment; *Commun. Statist.—Theor. Meth.* **1980,** *A9* (17), 1749–1842.
19. Morgan, M.G.; Henrion, M. *Uncertainty, A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*; Cambridge University Press: New York, 1990.

About the Authors

Michael Sohn is a scientist in the Indoor Environment Department at Lawrence Berkeley National Laboratory (LBNL). Pamela Reynolds and Navtej Singh were research interns at LBNL. Reynolds is now a graduate student in the Graduate School of Journalism at the University of California, Berkeley. Singh is now a mathematics lecturer at the University of Hawaii, Hilo. Ashok Gradgil is a senior staff scientist in the Indoor Environment Department at LBNL. Please direct correspondence to Michael Sohn, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop: 90R3058, Berkeley, CA 94720; e-mail: mdsohn@lbl.gov.

ERRATUM

"Rapidly Locating and Characterizing Pollutant Releases in Buildings," Sohn, M.D.; Reynolds, P.; Singh, N.; and Gadgil, A.J. *J. Air & Waste Manage. Assoc.* **2002**, *52*, 1422-1432: In Figures 4a through 4c (page 1428), the lines identified as "room 1" were mislabeled; they should be labeled "room 2" (see correct figure below). We are sorry for any confusion this might have caused the readers.

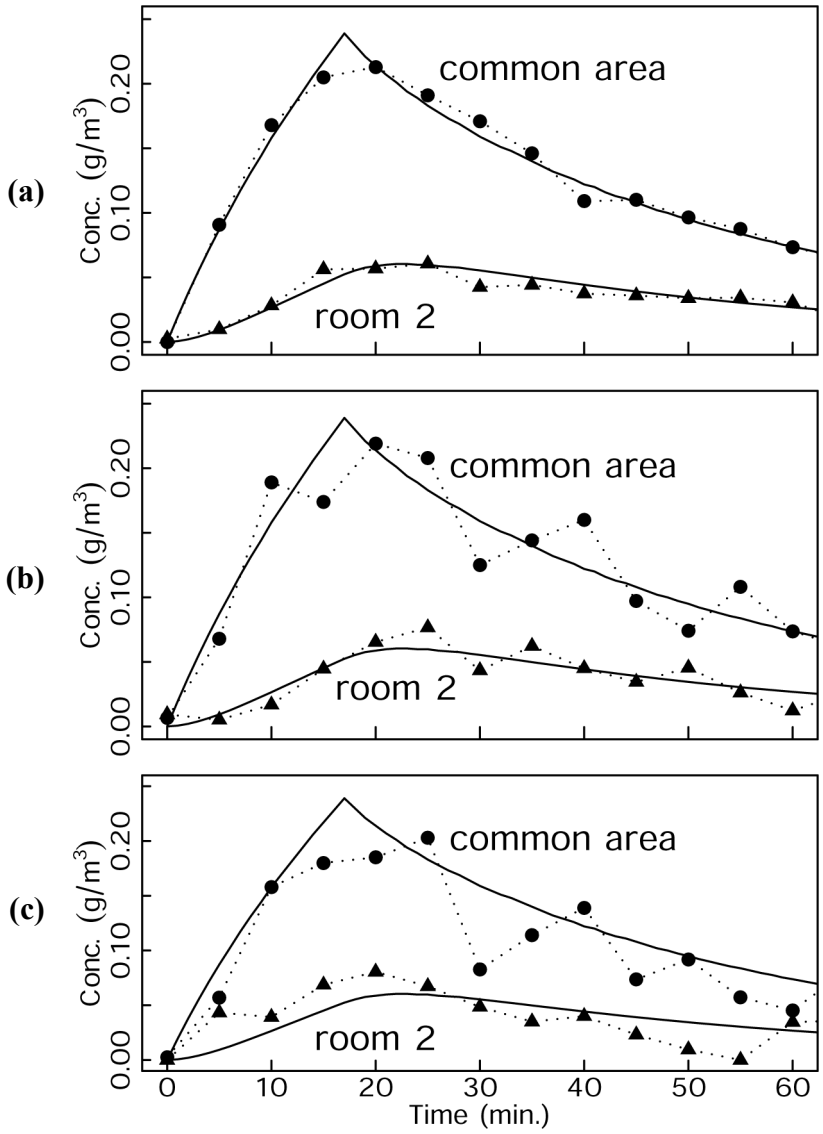

Figure 4. Synthetic data representing (a) high-, (b) medium-, and (c) low-quality measurements in the common area zone and room 2. The solid lines represent the model simulation originally used to generate the data.